American Journal of Computer Science and Technology
2022; 5(4): 210-216
http://www.sciencepublishinggroup.com/j/ajcst
doi: 10.11648/j.ajcst.20220504.13
ISSN: 2640-0111 (Print); ISSN: 2640-012X (Online)

Chromosome Segmentation Analysis Using Image Processing Techniques and Autoencoders

Amritha Pallavoor[1], Prajwal Anagani[1], Sundareshan Tambarahalli[2], Sreekanth Pallavoor[3, *]

[1]Department of Computer Science & Engineering, PES University, Bangalore, India

[2]Dr. Rao's Genetics Laboratory and Research Center, Bangalore, India

[3]CARAIO Technologies, Bangalore, India

Email address:

sreekanth@caraiotech.com (Sreekanth Pallavoor)

[*]Corresponding author

To cite this article:

Amritha Pallavoor, Prajwal Anagani, Sundareshan Tambarahalli, Sreekanth Pallavoor. Chromosome Segmentation Analysis Using Image Processing Techniques and Autoencoders. *American Journal of Computer Science and Technology*. Vol. 5, No. 4, 2022, pp. 210-216.
doi: 10.11648/j.ajcst.20220504.13

Received: October 14, 2022; **Accepted**: November 10, 2022; **Published**: November 23, 2022

Abstract: Chromosome analysis and identification from metaphase images is a critical part of cytogenetics based medical diagnosis. It is mainly used for identifying constitutional, prenatal and acquired abnormalities in the diagnosis of genetic diseases and disorders. The process of identification of chromosomes from metaphase is a tedious one and requires trained personnel and several hours to perform. Challenge exists especially in handling touching, overlapping and clustered chromosomes in metaphase images, which if not segmented properly would result in wrong classification. This study proposes a method to automate the process of detection and segmentation of chromosomes from a given metaphase image, and in using them to classify through a Deep CNN architecture to know the chromosome type. There are two methods to handle the separation of overlapping chromosomes found in metaphases - one method involving watershed algorithm followed by autoencoders and the other a method purely based on watershed algorithm. These methods involve a combination of automation and very minimal manual effort to perform the segmentation, which produces the output. The manual effort ensures that human intuition is taken into consideration, especially in handling touching, overlapping and cluster chromosomes. Upon segmentation, individual chromo- some images are then classified into their respective classes with 95.75% accuracy using a Deep CNN model. Further, a distribution strategy is imparted to classify these chromosomes from the given output (which typically could consist of 46 individual images in a normal scenario for human beings) into its individual classes with an accuracy of 98%. This study helps conclude that pure manual effort involved in chromosome segmentation can be automated to a very good level through image processing techniques to produce reliable and satisfying results.

Keywords: Chromosome Analysis, Karyotyping, Cytogenetics, Chromosome Segmentation, Autoencoder, Squeezenet, Watershed Algorithm

1. Introduction

Human Chromosome Analysis involves the identification of 23 pairs of Chromosomes in human cells, out of which the first 22 pairs are called Autosomes and the 23rd pair is the Sex Chromosome. This process is called Karyotyping, and it is typically performed by experts in the field of Cytogenetics. Its potential is immense in the early detection/diagnosis of diseases including Constitutional abnormalities, Prenatal and Acquired abnormalities. However, the whole process is very manual, labour intensive, and hence prone to human errors and fatigue, which could effectively result in delays in report generation and inaccurate/faulty reports.

The process of Karyotyping includes multiple stages, and the whole exercise involves a lot of manual effort to go through the images of chromosomes, look for abnormalities if any and confirm those and prepare the final report. The adoption of Artificial Intelligence and Image Processing is pertinent to this problem domain, where information exists in terms of digital images.

The focus of this paper is to develop an automated image segmentation process that identifies and segments the individual chromosomes (46 in number for normal cases) from the metaphase image obtained through a microscope, and further classify them with a deep learning model. This process, when implemented through a software, is intended to help freshers and students in cytogenetics learn to use their intuition and domain knowledge to separate overlapping chromosomes in an easier and effective way by combining human and machine intelligence, thereby making their training process much simplified. Additionally, it can function just as an aid in diagnosis as well.

Routine chromosome analysis in cytogenetics requires culturing of samples for a certain period depending upon the sample type, followed by metaphase slide preparation and image capture of metaphase using a microscope. This process involves one of the various staining methods, out of which this research considers the case involving Q-banding staining method and the G-banding staining method separately. The Q-banding staining method is a fluorescent staining method, which uses quinacrine, is used to identify individual chromosomes and their structural anomalies, whereas G-banding is done by using Giemsa or Leishman stains to produce thin, alternating bands along the length of the entire chromosome that create unique patterns through which identification is done. The characteristic banding pattern can be used to identify each chromosome accurately. These images are then loaded onto a software and manually analyzed to segment them using human expertise and through imparting visual cognitive effort. This research proposes to automate this part by implementing Image Processing techniques to segment chromosomes more efficiently, especially for early-stage users in the field. After segmentation of the chromosome images through this approach, these images are to be used with a Deep CNN architecture to train and achieve the automated classification of these chromosome images to a high accuracy level to produce the Karyogram output.

The contribution is 4 fold:

1) Identification and separation of individual chromosomes.
2) Detecting and separating touching, overlapping, and chromosome clusters.
3) Predicting the chromosome classes for all individual chromosome images also handling abnormality detection.
4) A distribution algorithm that increases overall accuracy.

2. Related Work

The analysis of a Karyotype has a wide range of uses in the cytogenetics field, its most important and common usage lies in prenatal diagnosis and genetic disease detection. The current industry method involves manual segmentation and classification by a trained person to classify and detect anomalies.

The methods and research in this paper for chromosome image segmentation is performed on G-Band and Q-Band metaphase images. Classification of chromosomes is constrained to Q- Band metaphase images as that's where better accuracy was achieved. Research is being continued for G-Band image classification to obtain the best possible accuracy.

N. Xie et al. [1] proposed the use of multiple input convolutional neural networks (CNN) and geometric optimization, called mCNN GO for classification. They used Mask R-CNN for the segmentation of individual chromosomes from the metaphase images and classified the sub-images using the mCNN GO. They also performed chromosome straightening with a medial axis locating algorithm, and achieved around 95.644% accuracy for segmentation.

M. S. Al-Kharraz et al. [2] proposed a segmentation method where individual chromosome detection was done using YOLOv2 CNN followed by some chromosome post-processing. This step achieved 0.84 mean IoU. They used VGG19 for further processing and classification and obtained an accuracy of 94.11% on the BioImLab Q-Band image dataset. They worked with metaphase images containing non-overlapped chromosomes.

By adding a number of layers onto the U-Net architecture, H. M. Saleh et al. performed overlapping chromosome semantic segmentation by implementing TTA and reached 99% accuracy [3].

H. A. Al-Ameri et al. [4] detected overlaps by performing thinning of the image using a Morphological operation. They found the cut points of the intersection by implementing an algorithm with a predefined 7x7 mask.

E. Poletti et al. [5] used features extraction methods, and extracted medial axis, polarisation and length of individual chromosomes, to then classify them using an MLP network, they also used a multi-stage decision tree to find the polarisation of the chromosome. They achieved an accuracy of 95.6%.

E. Poletti et al. [6] proposed using extensive features that were extracted from the chromosome images, and improved estimation of the medial axis. Feature re-scaling and normalising techniques take full advantage of the results of the polarisation step, re- ducing the intra-class and increasing the inter-class variances. They also use a rule-based approach that works on features to identify the polarisation. An MLP is used for classification and the accuracy obtained is 94%.

E. Poletti et al. [5, 6] employ a distribution strategy that takes into account the number of chromosomes constraint and helps them classify chromosomes from a karyotype with better accuracies.

M. Al-Kharraz et al. [7] proposed an ensemble of 3 Deep CNN models to further increase the classification accuracies. Their method achieves a classification accuracy of 97% with the ensemble of VGG19, ResNet50 and MobileNetv2.

S. Swati et al. [8] proposed a multistage architecture that includes a network that upscales the image to a higher resolution, using super-resolution layers to upscale. They further use an Xception or a ResNet50 to classify the scaled images. The highest accuracy was achieved by using Xception and it is 93%.

X. Liu et al. [9] proposed using a super-resolution net and self-attention negative feedback network and combining it with Deep CNNs to obtain a classification method (SRAS-net). The class imbalance of X and Y chromosomes was tackled by using SMOTE to generate additional Y samples. They achieved an accuracy of 97.5%.

M. Sharma et al. [16] focused on crowdsourcing, and enabling large segmented dataset which can be fed to DNN models for automatic classification. However, this is dependent on human effort a lot, while some automation is there. Human in the loop is a key consideration factor here.

R. L. Hu et al. [17] the distinction between partially overlapping chromosomes was done by using a neural network based image segmentation method. This was applied on a synthetic dataset by using U-Net for segmentation. The results achieved IoU scores of 94.7% for the overlapping region and 88-94% for the non-overlapping chromosome regions.

3. Dataset

Two sets of chromosome image datasets were used for the research and development of this project. The first dataset includes Q Band chromosome images which are obtained by staining chromosomes with Quinacrine, a fluorescent dye in the laboratory. It consists of 230 images of individual chromosomes for each type, across all 23 types of chromosome classes (including X and Y as class 23). The dataset also has 117 metaphase images from which these individual chromosome images are taken. These images include a wide range of chromosome orientations which are straight, bent and so on and also include touching, overlapping and clusters of chromosomes in the metaphase images. The individual chromosome images are all in the right polarity with the p-arm placed above the q-arm, and was used for the training of deep learning models to predict the chromosome classes. This is a publicly available dataset found online from the BioImLab- Laboratory of Biomedical Imaging [10].

The second dataset contains G-Band metaphase images which are prepared by staining chromosomes with the Giemsa stain. Here, the banding patterns are more pronounced and can be seen more clearly when compared to Q-Band stained images. This dataset was obtained from Dr. Rao's Genetics Laboratory and Research Center, Bangalore, India. The metaphase images also include a wide range of chromosomes that are touching, overlapping or clustered.

4. Methodology

The metaphase image comprises 46 chromosomes for a normal human cell or it may be 45 or 47 chromosomes typically in cells involving numerical abnormalities. These chromosomes are scattered in the image and they occur in ways that require complex segmentation methods. These are categorized into 4 main categories-

1) Individual isolated chromosomes
2) Touching chromosomes
3) Overlapping chromosomes
4) Clusters of chromosomes

A. Segmentation - Initial Stage

The G-Band metaphase images (Refer Figure 1) occur with pixel intensities varying from 0 to 255. It generally includes chromosomes in gray lying on a white background. Object detection is first performed to locate the chromosomes on the image. Noise reduction is done to the image by applying the medianBlur filter over the image. This helps smoothen the pixels inside the chromosomes and focuses only on the outermost contour of the chromosomes and not the bands found inside.

The Otsu thresholding [14] algorithm is used on the denoised grayscale image for binarization. This way, the background is separated from the foreground (chromosomes). It is important to implement adaptive thresholding to find the optimal value as this allows the algorithm to work on metaphase images of varying intensities and not just conform to a single threshold value. The next step involves the detection of chromosome outlines. One of the approaches for this is to use the Canny Edge detection [13] method. Upon using the Canny Edge algorithm on the metaphase image with the aperture size set to a value equal to 5, the boundaries of the chromosomes are clearly identified. Even finer details like the banding patterns inside the chromosome are visible to a great extent (Figure 2).

As the initial step is to detect and segment the individual chromosomes, the Morphological Gradient [15] is used to detect the external boundaries alone. The outlines of the chromosomes are found by applying the Morphological Gradient operator using a 2x2 kernel in the image. Morphological Gradient is essentially the difference between the dilated and eroded versions of the image which produces the outline of the boundaries (Figure 3). Contours are found on this image and by using it as a mask, each contour is iterated through to get the chromosome Region Of Interest (ROI). Once the ROI is obtained, the chromosome is placed on a three-channel white background image. This ensures that any noise, or parts of other chromosomes are absent in that particular image. The minimum area rectangle bounding box is calculated for that chromosome, and this way multiple single chromosome images are saved from the metaphase for each contour. At this point, individual chromosomes are fully segmented and preprocessed, and clusters and overlaps have completed the initial stage of object detection.

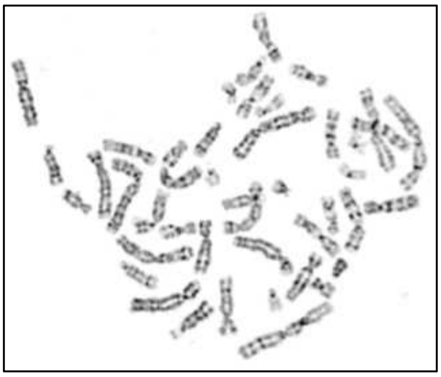

Figure 1. *Metaphase Image.*

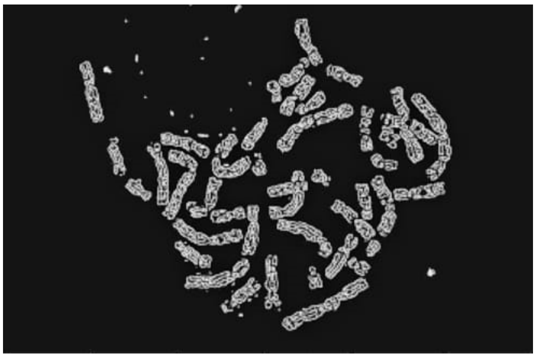

Figure 2. *After applying Canny Edge Detection.*

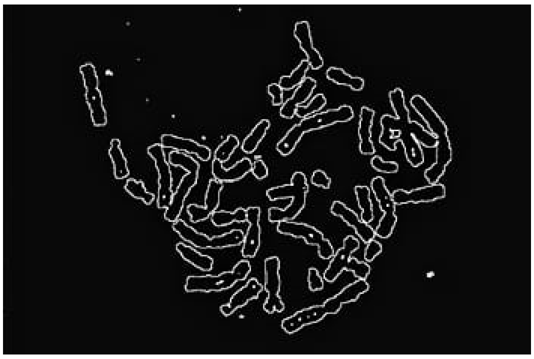

Figure 3. *After applying Morphological Gradient.*

B. Detection of Overlaps and Clusters

The next stage of the chromosome segmentation pipeline involves the detection of overlaps and clusters from the set of chromosome images that is received through the initial stage of segmentation. To achieve this, each sub-image is skeletonized after some noise reduction (Figure 4). The skeleton of an image is the eroded, thinned-down representation. The chromosomes are now shown as lines. If there exists an overlap, the skeleton lines intersect at a point. This strategy is used to detect chromosome overlaps. The same works in the case of touching and cluster chromosomes as well.

The presence of an intersection is found by locating a white pixel and obtaining its neighboring pixels. The neighboring pixels have a value of 1 if it's white or 0 if it's black. A list of such pixels surrounding the pixel of interest is obtained and compared with a predefined list of all possible pixel value combinations where an intersection exists. This algorithm returns the intersection coordinates if and when detected.

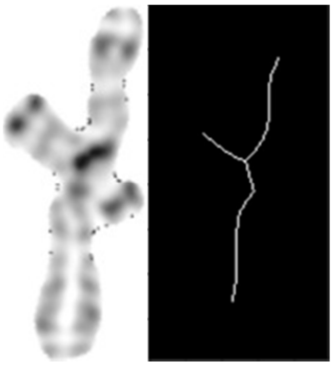

Figure 4. *Overlap and its skeleton.*

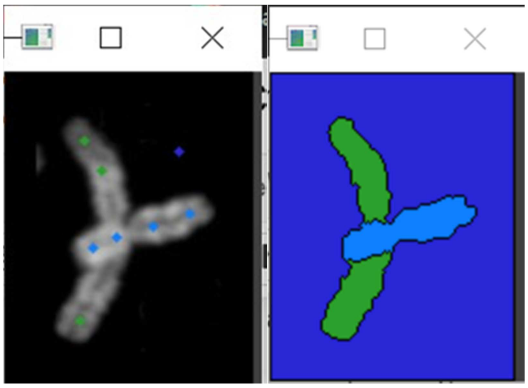

Figure 5. *Overlap image (left) and its segmented output (right).*

C. Segmentation - Separation of Overlapping Chromosomes

The research has led to the discovery of two types of approaches for solving this using Image Processing based techniques-

1. Watershed Algorithm is followed by Autoencoding to fill the missing intersection region that was initially covered by the overlap.
2. Watershed Algorithm that considers the overlap intersection as a part of both chromosomes after separation.

Method 1: This method works best with Q-Band chromosome images. Assume an image having two overlapping chromosomes. The watershed algorithm is used to segment the two chromosomes. Here, the markers for specifying seeds to the segmentation algorithm are provided manually. This ensures that the user has more control over how they perceive the particular chromosome overlap. A clinician would have to decide if the intersection is due to two chromosomes crossing each other, or due to two bent chromosomes touching at the bending point.

Once the seeds are placed appropriately at the regions on the image (Figure 5 - left), the segments are generated accordingly, and drawn as a new image with the segments shown correspondingly (Figure 5 - right). The regions are differentiated by changing the segment's color. All segments belonging to a particular chromosome have the same color.

Each segment is then placed on a new background image unique to each chromosome. The chromosome lying above includes the intersection region after segmentation, whereas the chromosome lying beneath has an empty region where the intersection initially was. To help fill this gap, an autoencoder was trained with the BioImLab chromosome image dataset. An image autoencoder is a neural network that learns to first decompose the data into smaller objects and then reconstruct the original image using it to match the original as closely as possible. Autoencoders are widely used to reconstruct missing data, smoothen images, or even reduce noise. The autoencoder model was trained using the Tensorflow framework and used it to fill the gaps after the watershed segmentation if any. It was also found that autoencoding the entire image resulted in loss of valuable information like the chromosome banding pattern which is extremely important for CNN models to classify.

Hence, only the missing region of the chromosome is autoencoded (Figure 6).

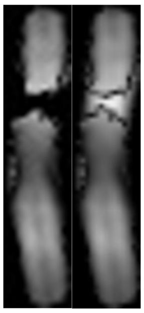

Figure 6. *Chromosome with a missing region (left) after autoencoding (right).*

This method involves a few drawbacks. Firstly, an image autoencoder model was able to be trained with good accuracy only for the Q-Band chromosome dataset. The autoencoder for the G-Band chromosomes did not perform with good accuracy. Secondly, the autoencoded image did not get classified correctly by the Deep CNN model at all times.

To resolve the above drawbacks, Method 2 is proposed.

Method 2: This method supports G-Band and Q-Band chromosomes. The Watershed Algorithm described in Method 1 is implemented here as well. The seeding of chromosome segments is done in such a way that the intersection region of the overlap is specified as a separate segment from the rest of the chromosome segments (Figure 7 - right). Finally while integrating the chromosome segments, the intersection region is integrated in both of the final separated chromosome images (Figure 8).

D. Classification

After completion of the segmentation of chromosomes from the initial metaphase image, the result includes 46 individual chromosome images (assuming a case with no abnormalities). The next step is to correctly classify the chromosomes into their respective pair numbers - there are a total of 23 pairs, with the last pair including the sex chromosomes which is X and X in the case of females or X and Y in case of males. To do the classification, a Deep Convolutional Neural Network model - SqueezeNet [11] was trained. The SqueezeNet architecture is designed to reduce the number of parameters by "squeezing" parameters with the help of fire modules that use 1x1 convolutions. This results in a 50x reduction in model size when compared to AlexNet [12], while maintaining good prediction accuracy.

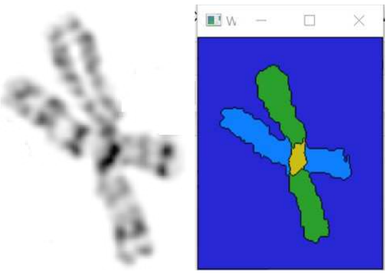

Figure 7. *Overlap (left) segmented with the intersection seeded separately (right).*

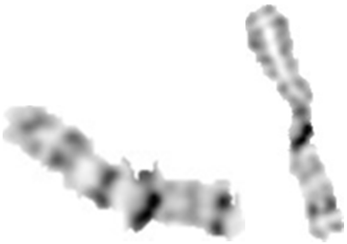

Figure 8. *Chromosomes after separation from the overlap.*

The pair number for every segmented chromosome image is then predicted by using the trained CNN model, with good accuracy.

E. Distribution Algorithm

To further improve the final classification accuracy, the distribution algorithm is used. The idea is to take advantage of the constraint on the number of chromosomes for a given metaphase and redistribute erring predictions to best classify the chromosomes.

Steps-

1) Get individual predictions for each pair.
2) Find pairs that are lacking (less than 2 chromosomes are predicted for it).
3) For each of the lacking pairs, search crowded pairs (more than 2 chromosomes are predicted for it) for an image that has the highest score for the class that the lacking pair is in.
4) Reassign that image to the lacking pair.
5) Repeat until there are no lacking pairs left.

5. Results

The segmentation process which involves the detection and separation of purely individual chromosomes from the metaphase image works efficiently and produces approximately 95% accuracy.

In cases of segmentation involving overlap detection and separation of chromosomes, the proposed algorithm works with most images and produces around 94% accuracy after performing tests on multiple image segments.

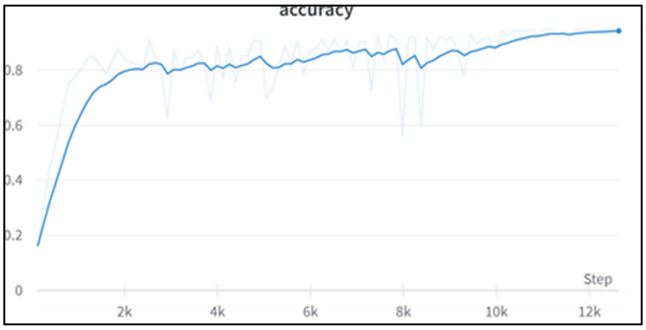

Figure 9. *Validation Accuracy of 95.75% - SqueezeNet model.*

The SqueezeNet model trained on all 23 pairs of chromosome images produced an accuracy of 95.75%. The model was prepared by using an early stopping callback to prevent overfitting of the data. It trained for 119 epochs using the K-fold cross-validation strategy (5 folds). Refer Figure 9 and Figure 10 for visualization of the metrics. Using the distribution algorithm on the model predictions pushed the accuracy up to 98% when the segmented images from the metaphase are given as the input to the SqueezeNet model for classification. This end-to-end segmentation-classification pipeline thus validates the proposed approach towards automating the karyotyping process.

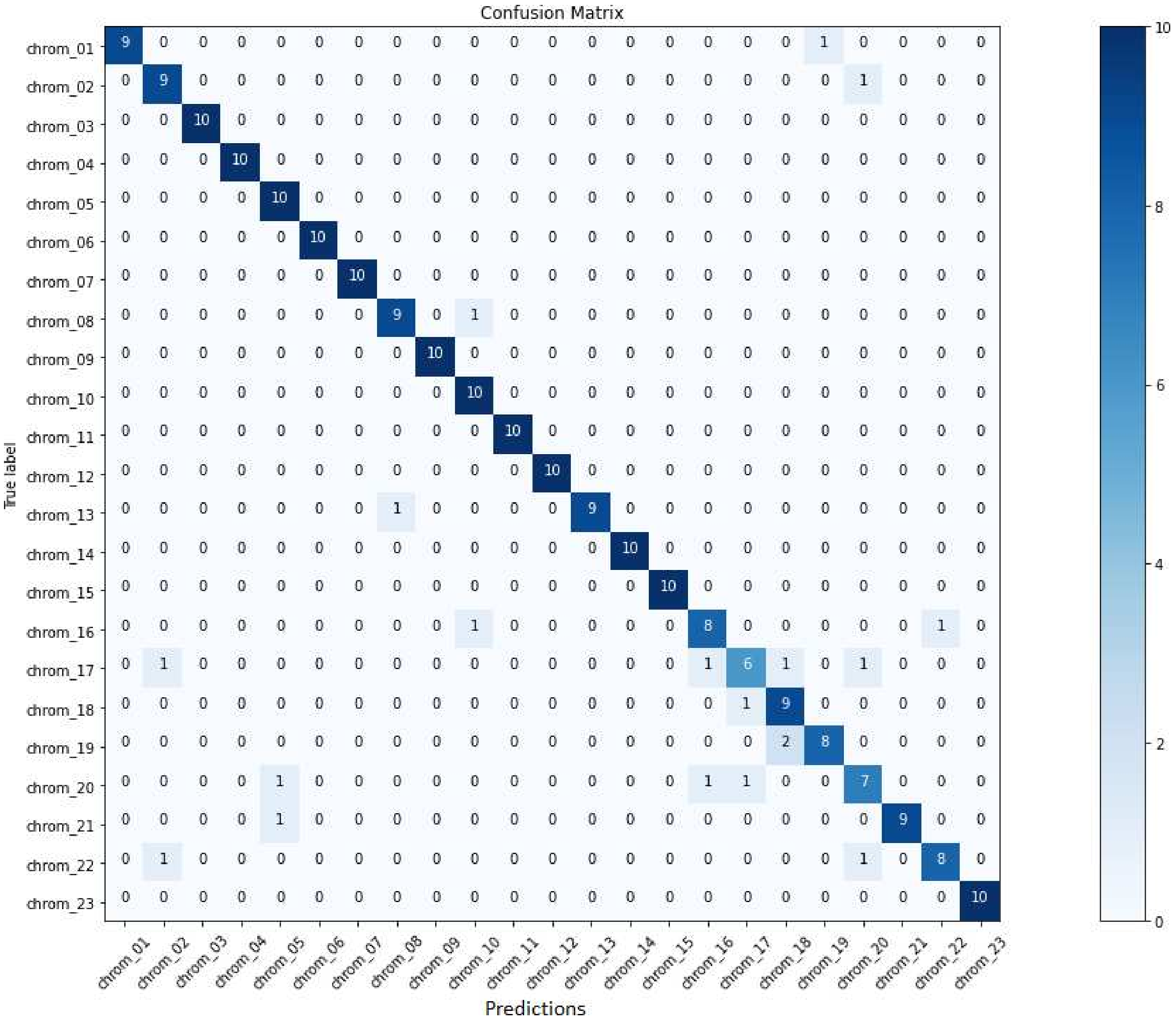

Figure 10. *Confusion matrix for the test data predictions.*

6. Conclusion

Karyotyping process is predominantly a manual effort driven task today and is generally very tedious. The main aim of this research was to simplify the segmentation of chromosomes in metaphase images, so that in particular complexities involved in the separation of touching, overlapping, and cluster chromosomes are resolved through combination of automation and very minimal manual efforts. This helps as an aid and augmentation to cytogeneticists in the task and help freshers in the field to get trained and learn the domain better and faster. Through this research, efficient methods to initially segment the individual chromosomes were found, and future work could involve the use of the Canny Edge detection algorithm to further detect internal banding patterns found in chromosomes.

This research proposes a method that performs the overlap separation by using the watershed algorithm and segments the chromosome regions by manual seeding followed by techniques like contouring and image cropping. One of the approaches also involved the use of image autoencoding. Further, the chromosome class prediction accuracy after segmenting individual chromosomes from the metaphase image was validated. This was done by training a Deep CNN - SqueezeNet model that produced an accuracy of 95.75% on the dataset. By employing the distribution algorithm, the accuracy was enhanced to 98%.

This way, the process of handling metaphase images consisting of touching, overlapping, and clustered chromosomes in real-world can produce more accurate segmentation output. Those output images can then be fed into various classification algorithms for much better prediction outcomes of chromosome classes with higher accuracies. The impact of this overall, will result in automation of the karyotyping process to predict chromosomal defects with high accuracy.

References

[1] N. Xie, X. Li, K. Li, Y. Yang and H. T. Shen, ”Statistical Karyotype Analysis Using CNN and Geometric Optimization,” in IEEE Access, vol. 7, pp. 179445-179453, 2019, doi: 10.1109/ACCESS.2019.2951723.

[2] M. S. Al-Kharraz, L. A. Elrefaei and M. A. Fadel, ”Automated System for Chromosome Karyotyping to Recognize the Most Common Numerical Abnormalities Using Deep Learning,” in IEEE Access, vol. 8, pp. 157727-157747, 2020, doi: 10.1109/ACCESS.2020.3019937.

[3] H. M. Saleh, N. H. Saad, N. A. Mat Isa, ”Overlapping Chromosome Segmentation Using UNET: Convolutional Networks with Test Time Augmentation ” in Procedia Computer Science, Vol 159, pp. 524-533, 2019.

[4] H. A. Al-Ameri and W. Al-Hameed, ”New algorithm for separation overlapping touching chromosomes”, 2020 J. Phys.: Conf. Ser. 1530.

[5] E. Poletti, E. Grisan, A. Ruggeri, ”Automatic classification of chromo- somes in Q-band images” in Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, 1911–1914.

[6] E. Poletti, E. Grisan, A. Ruggeri, ”A modular framework for the automatic classification of chromosomes in Q-band images”, Computer Methods and Programs in Biomedicine, Vol 105, pp. 120-130, 2012.

[7] M. Al-Kharraz, L. A. Elrefaei, M. Fadel, “Classifying Chromosome Images Using Ensemble Convolutional Neural Networks” in Applications of Artificial Intelligence in Engineering, pp. 751, 2021.

[8] S. Swati, M. Sharma and L. Vig, ”Automatic Classification of Low- Resolution Chromosomal Images”, in Computer Vision – ECCV 2018 Workshops, Vol 11134, pp. 315-325, 2019.

[9] X. Liu, L. Fu, J. Chun-Wei Lin, S. Liu ”SRAS-net: low-resolution chromosome image classification based on deep learning” in IET Systems Biology, 16 (3-4), 85–97 (2022). https://doi.org/10.1049/syb2.12042

[10] E. Grisan, E. Poletti, A. Ruggeri, ”Automatic segmentation and disen- tangling of chromosome in Q-band prometaphase images”, IEEE Trans Inf Technol B, 2009.

[11] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer, ”Squeezenet: AlexNet-level accuracy with 50x fewer parameters and ¡0.5MB model size”, arXiv: 1602.07360, Nov 2016.

[12] A. Krizhevsky, I. Sutskever, G. E. Hinton, ”ImageNet Classification with Deep Convolutional Neural Networks” in Communications of the ACM, Vol 60, Issue 6, June 2017 pp 84–90, https://doi.org/10.1145/3065386

[13] J. Canny, ”A Computational Approach to Edge Detection,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI- 8, no. 6, pp. 679-698, Nov. 1986, doi: 10.1109/TPAMI.1986.4767851.

[14] N. Otsu, ”A Threshold Selection Method from Gray-Level Histograms,” in IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

[15] J. Rivest, P. Soille, S. Beucher, ”Morphological gradients” in Journal of Electronic Imaging, Oct 1993, 2 (4): 326-336, doi: 10.1117/12.159642

[16] M. Sharma, O. Saha, A. Sriraman, R. Hebbalaguppe, L. Vig and S. Karande, ”Crowdsourcing for Chromosome Segmentation and Deep Classification,” 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 786-793, doi: 10.1109/CVPRW.2017.109.

[17] R. L. Hu, J. Karnowski, R. Fadely, J. Pommier, ”Image Segmentation to Distinguish Between Overlapping Human Chromosomes”, 31st Conference on Neural Information Processing Systems (NIPS 2017), arXiv: 1712.07639v1.